DOCUMENT RESUME

ED 205 550 TM 810 429

AUTHOR Koch, William R.
TITLE Parameter Invariance of the Graded Response Latent Trait Model.
PUB DATE Apr 81
NOTE 25p.: Paper presented at the Annual Meeting of the National Council on Measurement in Education (Los Angeles, CA, April 14-16, 1981).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Attitude Measures: Correlation: Difficulty Level: *Goodness of Fit: Knowledge Level: *Latent Trait Theory: Maximum Likelihood Statistics: Regression (Statistics): Simulation: Test Items
IDENTIFIERS Estimation (Mathematics): *Invariance Principle: Likert Scales: Polychotomous Scoring: *Response Model: Two Parameter Model

ABSTRACT

The two-parameter graded response latent trait model was applied under various conditions to two simulation data sets and to data obtained from a Likert-type attitude scale. The purpose was to investigate the invariance property of the item and person parameter estimates for polychotomously scored data. Correlation and regression analyses, as well as standardized difference scores, were used to compare corresponding sets of parameter estimates for high-low ability groupings of people and for easy-difficult splits of items. In general, the results successfully demonstrated the parameter invariance of the graded response model. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
W. R. Koch
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED205550

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
■ The document has been reproduced as received from the person or organization originating it.
□ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

TM 810 429

PARAMETER INVARIANCE OF THE GRADED RESPONSE LATENT TRAIT MODEL

William R. Koch
University of Texas at Austin

It is well known that the traditional methods of classical test theory yield sample-dependent results. For example, the reported reliability, the standard error of measurement, the difficulty and discrimination levels of the items, and the raw or standard scores of a test all depend on the particular administration of the test from which these indicators are derived. The problems resulting from this dependence between a measuring device and the persons being measured have, for years, led measurement specialists in the behavioral sciences to envy their counterparts in the physical sciences where, for instance, a gram is a gram, regardless of the particular object being weighed.

Therefore, it is not surprising to witness the enthusiasm (and skepticism) with which researchers in education and psychology have received the claims of the invariance (Lord and Novick, 1968; Lord, 1980) or objectivity (Wright, 1968) properties of the latent trait and Rasch (1960) test models, respectively. Given that the item response data fit the assumptions of the particular latent trait model being considered, it is possible theoretically to perform measurement in which an independence exists between the measuring instrument and the persons whose traits are being measured.

In practice, this property means that any subset from a set of items that all measure the same trait may be used to

Paper presented at the Annual Meeting of the National Council on Measurement in Education, Los Angeles, 1981.

measure persons on the same scale. Thus, a person will receive the same score (trait estimate) regardless of whether the test is an easy or difficult one, or a long or short one. The accuracy of such estimates, however, may differ considerably. On the other side of the coin, any subsample of persons may be used to obtain item parameter estimates (e.g. difficulty levels) on a common scale. Thus, items will receive the same difficulty parameter estimate regardless of whether the sample used for the calibration has a high or low ability level.

It is important to note, though, that these parameter estimates are only invariant within a linear transformation. For example, the ability estimates that result from persons taking two different tests measuring the same trait will not be exactly the same. Rather, they will differ by a constant amount, because the ability scales will have different origins but will share the same unit of measurement. The same is true of item difficulty parameter estimates. However, item discrimination parameter estimates obtained from two groups of people will have different units of measurement but will share a common origin. Simple linear regression is, therefore, an appropriate and straightforward means of assessing the property of parameter estimate invariance.

Several studies in the literature have successfully demonstrated the objectivity or invariance property of the Rasch one-parameter logistic model and of the three-parameter logistic model. Indeed, the existence of the property is usually assumed as the basis for the many applications that

are made of latent trait theory to practical testing problems. Studies of item bias, tailored testing, test equating, item linking to form item pools, information functions, and relative efficiency of tests are all based to varying degrees on the invariance property of the parameters in latent trait models. While some of these applications are merely easier to perform with latent trait theory than with classical testing theory, others cannot be done with classical testing theory. In this regard, it is worth noting that the use of item proportion correct as a function of observed score, using number-right scores, or even estimated true scores preclude the possibility of an independence of the measuring instrument from the persons being measured.

Despite the numerous advantages offered by latent trait theory over classical testing theory, a fairly serious limitation is present when applications are restricted to the binary response data required by the Rasch and the three-parameter logistic models. There are a great many applications of the methods of classical testing theory to situations where dichotomously scored items are not appropriate. For example, it is quite common to score problem solving tests (whether in mathematics, engineering, chemistry, physics, etc.) according to the degree to which the complete solution is obtained. Often the problems are broken down into a sequence of steps that must be solved hierarchically to get the correct answer.

Other examples of polychotomously scored instruments include attitude scales, personality trait measures, interest

inventories, and essay or short answer tests. If latent trait methods are to become widely accepted and practiced, they must be generalized from the current applications that predominantly are restricted to binary scoring. Fortunately, such models as the nominal, graded, and continuous response latent trait models have been developed. However, they have rarely been applied or investigated empirically.

The Graded Response Model

Given the fact that researchers have until recently shied away from the binary latent trait models due to their perceived mathematical complexity, it is easy to understand why the graded response model has largely been ignored (being even more complex). What follows is a brief tutorial to demonstrate that one may view the graded response model as a conceptually straightforward generalization of a binary response model.

In the binary model, we are interested in only two categories of responses, either correct or incorrect. Typically we plot an item response curve as the probability of a correct response conditional on the level of the latent trait being measured. Because all well-behaved probabilities range from 0.0 to 1.0, we can always obtain the probability of an incorrect response at each trait level by subtracting the probability of a correct response from 1.0. In subsequently attempting to estimate a person's trait level, we observe the person's response string of 1's and 0's across the items on the instrument. Using the assumption of local independence, the product of the probabilities

of the individual item responses yields the likelihood function of that particular response string. Because the likelihood is conditional on the trait level, the maximum of the function is the maximum likelihood trait estimate for the person. Note that our basic information needed was the probability of a response in either of the two categories.

Now the graded response model (Samejima, 1969) is appropriate for the case where there are more than two categories of possible responses. Furthermore, we assume that it is possible to rank order the categories according to some logical plan that relates high or low categories of response to high or low trait levels in the persons being measured. However, the basic information that will still be needed is the probability of a response in each of the various categories, conditional on the trait level.

For the sake of an example, we will consider the case of an item that may be scored on a scale from 1 to 5, meaning that there are five possible categories of response. Therefore, we will have four cutting points or boundaries that separate the response categories. Just as in the binary case, we will use item response curves to describe the conditional probability of responding in any particular category, but in a two-stage process. First, we find the function for a response in the first category versus a higher category (1 vs. 2,3,4, or 5), then the function for a response in the first or second category versus a higher category (1 or 2 vs. 3,4, or 5), and so forth until the last function for a response in the first four versus

category five (1,2,3, or 4 vs. 5). In total, there will be four separate response functions because there are five response categories. The second stage is simply to subtract the successive response functions from each other to obtain the desired probabilities of responses in each of the five categories separately. This procedure is illustrated in Figures 1-A and 1-B below. Figure 1-A shows the four response functions for a fictitious item, while Figure 1-B shows the separate conditional probabilities of responses in each of the five categories. Note that the probabilities for the extreme categories (1 and 5) are obtained by subtracting the response functions from 1.0 and 0.0, respectively, since there are no other response functions to use either below or above the extremes.

Figure 1-A
Category Response Functions.

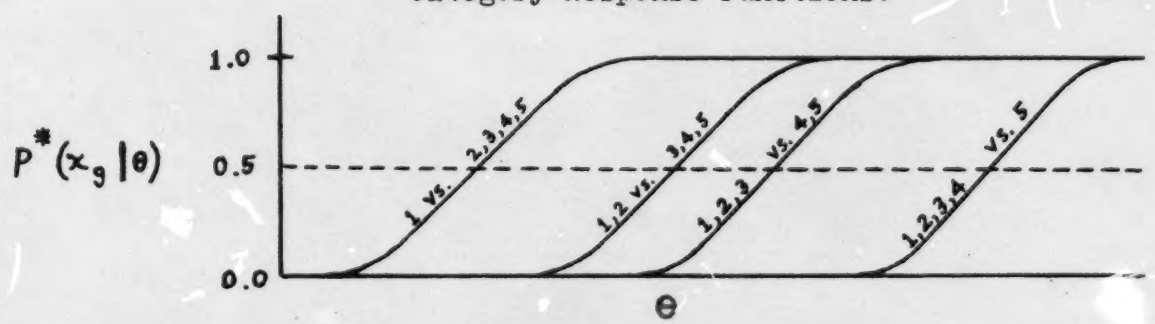
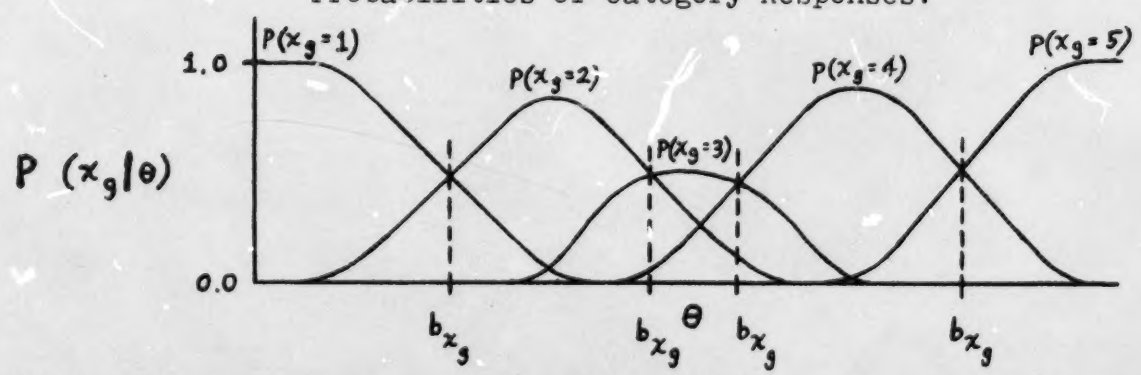


Figure 1-B
Probabilities of Category Responses.



In using the binary response models, we obtain a difficulty parameter estimate for each item by finding the point of inflection of the item response curve and projecting down to where it intersects with the trait scale (θ). In similar fashion, the category boundaries or cutting points (difficulties) for the graded response model are obtained by finding the point of inflection for each of the category response functions. For the item shown in Figure 1-A, this point is where the probability is at the 0.5 level. Projecting down to the θ scale yields the category boundaries (b_{x_g}) that are shown in Figure 1-B. Note also that the boundaries coincide with the points where the probability functions for the adjacent response categories intersect each other. Thus, for a five category item, we have four "difficulty" parameter estimates.

Samejima (1969) introduced the graded response model and presented the theory, rationale, and formula derivations for its use. The provision was made for each item on an instrument to vary in its discrimination index, so that each item had a set of category boundary estimates as well as <u>one</u> discrimination parameter estimate that served across all response categories for that item. Later, Andersen (1977) and Andrich (1978) developed ordered response models along similar lines, except that all items were assumed to have equal discrimination parameter estimates.

Computer programs are available to perform the item and person parameter estimation needed to use these models. The program by Kolakowski and Bock (1973) called LOGOG was designed

for both the graded response and the nominal response models. Also, by deleting a single program statement, the user can change from a model that allows the item discrimination estimates to vary to a model that holds them constant. Bock (1980) has reported that David Thissen at the University of Kansas has developed a version that is more efficient and is more stable in parameter estimation. Also, Ben Wright has begun to distribute a program for the ordered response model, where the items have equal discrimination parameters. Due to the ways in which Likert scales are typically constructed, it is often quite reasonable to assume that all items are equally discriminating.

METHOD

Instrumentation and Samples

Two simulation data sets were generated to investigate the parameter invariance of the graded response model. Each included simulated responses by 1,000 persons to instruments having 30 items, each of which was scored on a scale from 1 to 5. Both data sets were generated according to a simulation procedure that forced the response data to match a pre-specified factor loading matrix (Wherry, Naylor, Wherry, and Fallis, 1965). In the first case, a relatively pure one-factor data set was formed by setting the loading of each item to be .90, thereby insuring that the unidimensionality assumption of the latent trait model would be met. In the second case, one dominant factor was present, along with five smaller, specific factors.

Each item had a loading of .70 on the dominant factor and a loading of .60 that was randomly distributed to one of the five specific factors, with .00 loadings on the remaining four specific factors. The purpose was to simulate data that would approximate responses to a typical, high quality Likert scale. For convenience, the two simulation data sets will be designated as SD-U (simulation data, unidimensional) and SD-DS (simulation data, dominant and specifics).

The real data were the responses from a total sample of 491 teachers to a 40-item scale called the Audit of Administrator Communication (Valentine, 1978). The instrument was designed to measure the communication skills of school administrators with the teachers working under their supervision. Again, there were five possible categories of response to each item. The responses from the teachers at ten different school systems who rated their principals were collapsed for analysis. For convenience, the real data will be referred to as ADCOM.

Procedure

The LOGOG computer program (Kolakowski and Bock, 1973) was used to perform the latent trait item calibration and person parameter estimation for all three data sets. The output of the program included a trait parameter estimate for each person, a discrimination parameter estimate for each item, and four category boundary parameter estimates (difficulties) for each item. Also, the program provided chi-square tests of fit of the response data to the graded response model, both for each item and for the entire scale. Because it was desired to shorten

the 40-item ADCOM scale somewhat, the values of the chi-square fit statistics were used to successively delete 10 items that had significant lack of fit.

The method used to investigate the parameter invariance or objectivity of the graded response model was the same as has often been used by others (Wright, 1968; Whitely and Dawis, 1974) for the Rasch model. In order to study the invariance of the item parameter estimates, the total number of examinees was divided into a low score and a high score group on the basis of total scale summated raw scores. The raw score corresponding to the fiftieth percentile was used as the cutting point. This was done separately for each of the three data sets. Item parameters were then estimated using LOGOG with separate runs for each subgroup, so that six sets of item parameter estimates were obtained. According to the latent trait theory, it was expected that these estimates for the low and high score groups would differ only through a linear transformation.

To determine whether or not the invariance property of the person parameter estimates was demonstrated, two subscales of the "easiest" and "hardest" items were formed within each of the three instruments. The category boundary estimates (analagous to item difficulty parameters in the binary response models) were used as the criteria for the division of the items into subsets. In the context of the ADCOM attitude scale, these subsets should be thought of as consisting of items that tended to be easy or hard to agree with, respectively. Using the item parameter estimates obtained from the initial full scale calibrations,

maximum likelihood trait estimates were then calculated on both the easy and hard subscales for each person using LOGOG. Thus, within each of the three data sets, each person had two trait estimates. Again, it was expected that the resulting person trait estimates would remain invariant for each person across the easy and hard subscales.

To evaluate the invariance of the item parameter estimates for high and low score groups, correlation and regression analyses were performed. To measure the invariance of the person trait estimates, only correlation analyses were run because the LOGOG program put them on the same measurement scale automatically during trait estimation. Therefore, it was predicted that the means and standard deviations of the person trait distributions would be 0.0 and 1.0, respectively, with high correlations between the easy and hard trait estimates across the persons within each data set. Furthermore, to more closely assess the person trait invariance for the ADCOM real data, standardized difference scores (Wright, 1968) were computed. These difference scores were expected to be normally distributed standard scores if the only difference between the trait estimates for the easy and hard subscales was due to random measurement error.

It should be noted that maximum likelihood parameter estimation will not converge for any latent trait model in cases where persons get all the items correct or all incorrect, or in the present context, where persons have all 1's or all 5's in their response strings. Normally, this is not a frequent

occurrence, but forming subscales of easy and hard items tends to accentuate the problem. In regard to item parameter estimates, again one cannot obtain maximum likelihood category boundary estimates when there are no responses to a particular category, or when all the responses are in only one category. This situation usually does not occur, but it may be a problem when persons are divided into groups of low and high trait levels. The practical result of the method used by the present paper to study invariance, therefore, was that substantial numbers of persons and items dropped out of the analyses.

RESULTS

Item Parameter Invariance

The results for the invariance of the item parameter estimates are shown in Table 1. The first half of the table reports descriptive statistics of the item parameter estimates for both the low and high score groups over all three data sets. As one would expect, the items were more difficult for the low score groups than for the high score groups. For the item discrimination parameter estimates, both the means and standard deviations were fairly similar for the low and high score groups in each data set.

In the bottom half of the table are presented the data to indicate the degree of linear relationship between the item parameter estimates for the low and high score groups. For all three data sets the correlations between corresponding difficulty estimates were in the upper +.90's. Moreover, the regression

equations relating the difficulty estimates for the low and high score groups indicated that they differed by a constant, since the slopes were close to unity for all three data sets.

Insert Table 1 about here.

The results for the invariance of the item discrimination parameter estimates for the low and high score groups were not nearly as impressive as the difficulty parameter invariance. However, the data sets in the present study really did not provide for a fair test of the discrimination parameter invariance because of the extreme restrictions in the ranges of these discrimination parameters. For the SD-U and the SD-DS data sets, the discrimination parameters were preset as constants across all of the items, while the ADCOM scale had been constructed through item analysis to insure uniformly high discrimination values for the items (using traditional item score vs. total score correlations). Even still, the nine common items remaining in the ADCOM analysis across the low and high score groups had a correlation of r=+.71 for the discrimination estimates, and the intercept did not differ significantly from 0.0, as predicted.

Person Trait Parameter Invariance

Table 2 presents the results of the person parameter estimate invariance for the two simulated data sets, SD-U and SD-DS. As can be seen in both data sets, the raw score mean for the easy subscale was nearly twice the raw score mean for the difficult subscale, as would be expected. However, for the SD-DS scale, the person trait estimates from the easy and hard subscales had

Insert Table 2 about here.

virtually the same means and standard deviations, being close to their expected values of .00 and 1.00. Furthermore, the correlation of person trait estimates across the easy and hard subscales was r=+.87.

Unfortunately, the results for the SD-U data were distorted due to nonconvergence within the LOGOG estimation procedure for the difficult subscale, where three-fourths of the cases could not be estimated. Even so, for the 70 cases having trait estimates on both the easy and hard subscales, the correlation was r=+.60.

Insert Table 3 about here.

The results for the ADJOM scale are presented in Table 3. For the traditional raw scores, both the means and standard deviations were quite different for the easy and hard subscales. However, the latent trait person estimates were very similar, with the means and standard deviations being nearly identical. Also, the correlation between the corresponding trait estimates across the easy and hard subscales was r=+.71. Again, however, the sample size was reduced substantially due to nonconvergence of the LOGOG estimation. In terms of standardized difference scores, the mean and standard deviation were .05 and 1.48, respectively, which were somewhat higher than their predicted values of .00 and 1.00.

Because of the high frequency of nonconvergence of the LOGOG trait estimation across all three data sets, it was decided to rerun the tests of person parameter invariance using an empirical maximum likelihood estimation procedure. The results are shown in Table 4. For all three data sets, the

means and standard deviations of the person trait estimates were close to 0.0 and 1.0, respectively, regardless of whether the easy or difficult subscale was used. In addition, the means of the differences between the trait estimates from the easy and difficult subscales were very close to 0.0. Finally, the correlations between the easy and hard subscale trait estimates were all quite high: ADCOM, r=+.83; SD-U, r=+.92; and SD-DS, r=+.90. Some nonconvergence of trait estimation was encountered, but it was minimal compared to the high frequency of the LOGOG nonconvergence.

Insert Table 4 about here.

DISCUSSION AND CONCLUSION

Item Parameter Invariance

When the total sample of respondents in each data set was divided into high and low score groups, the results of the separate item calibrations generally supported the claim of an invariance property for the graded response model item parameter estimates. Certainly this was true for the category boundary parameter estimates, where the estimates differed by a linear transformation within each of the three data sets. However, the estimation was based on only one-third to one-half of the items on each scale due to the failure of the LOGOG program to provide reasonable estimates when the items were of inappropriate difficulty for the high or low score groups. In such cases, when it was even possible to converge on an estimate, the standard errors of estimation for the parameters became extremely high.

Although the item discrimination estimates from the ADCOM scale exhibited the invariance property to some degree, the data sets in the present study were not properly set up to adequately test the invariance property. This was due to the fact that the ranges of the item discriminations within each data set were extremely restricted. Therefore, even if the invariance property held up, the correlations would be near zero.

Person Trait Parameter Invariance

When the scales for each data set were divided into easy and hard subscales to obtain two separate trait estimates for each person, the results generally supported the claim of invariance for the latent trait graded response model. The problem of nonconvergence of the LOGOG maximum likelihood trait estimation clouded the results somewhat. But when the analyses were redone using an empirical maximum likelihood method (Reckase, 1974), the respondents received approximately the same latent trait estimates regardless of whether the estimates were derived from the easy or the difficult subscales.

It is not claimed that the methods used in the present study to investigate the invariance of the parameter estimates are definitive, or even as severe as might be encountered in practice. Rather, the study was a preliminary one to establish a basis for subsequent, more stringent tests of the invariance property. If reasonably good results had not been obtained here, one might be pessimistic about the possibility of forming large item pools of graded response items by means of linking together item parameters estimated from different samples of

persons, or pessimistic about using tailored testing or equivalent forms to estimate trait levels on the same scale regardless of the particular items administered. Certainly there is a need to take many of the latent trait techniques already shown to be of great benefit for the binary models and apply them to similar situations where items may be scored polychotomously. There is no reason why measurement specialists should not try to assess the item bias of attitude scale or personality trait items for different groups, or why the information provided by a scale should not be used to construct the optimal instrument for a particular application. As is currently being done in mastery testing, why should we not be able to measure changes in attitudes or personality traits on a fixed scale? In short, it is argued that the time has come to begin applying latent trait methods to measurement problems other than ability or achievement testing to see if similar advantages and benefits will accrue.

Table 1

Descriptive Statistics of LOGOG Item Parameter Estimates for High and Low Attitude Groups

Variable	Data Sets					
	ADCOM		SD-U		SD-DS	
	Low Group	High Group	Low Group	High Group	Low Group	High Group
Mean Difficulty	.50	-1.72	1.16	-4.08	1.18	-1.51
S.D. of Difficulty	2.64	2.36	3.18	2.67	2.18	2.15
Mean Discrimination	.95	.67	1.89	1.71	.80	.96
S.D. of Discrimination	.28	.18	.22	.23	.12	.12
Number of Items[a]	9	9	9	9	14	14
Number of Respondents	231	246	493	493	465	509

Relations Between Corresponding LOGOG Item Parameters for High and Low Attitude Groups

Data Sets	Difficulty Parameters		Discrimination Parameters	
	Correlation	Regression	Correlation	Regression
ADCOM	.97(36)	HI = -2.15 + .86 LOW	.71(9)	HI = .24 + .46 LOW
SD-U	.97(36)	HI = -5.03 + .81 LOW	.06(9)	———
SD-DS	.99(56)	HI = -2.66 + .98 LOW	.38(14)	———

Note. The numbers in parentheses indicate the number of parameters on which the correlations were based.

[a]The LOGOG program was unable to accurately estimate the item parameters for numerous items in all three data sets.

Table 2

Comparison of Person Attitude Estimates for Easy and Difficult Subscales of SD-U and SD-DS Data

SD-U Scale					
Subscale	Number of Items	Raw Score Mean	Raw Score Std. Dev.	LOGOG Est. Mean	LOGOG Est. Std. Dev.
Easy	12	45.42(995)	12.21(995)	-.01(730)	.90(730)
Difficult	12	25.70(995)	11.74(995)	1.35(131)	.48(131)

SD-DS Scale					
Subscale	Number of Items	Raw Score Mean	Raw Score Std. Dev.	LOGOG Est. Mean	LOGOG Est. Std. Dev
Easy	12	45.80(999)	9.32(999)	.03(936)	.99(936)
Difficult	12	26.28(999)	9.66(999)	.01(604)	.97(604)

Note. The correlations between the LOGOG attitude estimates for the easy and hard subscales for SD-U and SD-DS were r = +.60 (n = 70) and r = +.87 (n = 573), respectively. The parentheses in the body of the table indicate the number of respondents upon which the statistics were based. In all cases these sample sizes were less than 1,000 due to the deletion of cases of nonconvergence of the maximum likelihood attitude estimation and the deletion of cases with perfect scores.

Table 3

Comparison of Person Attitude Estimates for Easy and Difficult Subscales of ADCOM

Raw Score Attitude Estimates[a]				
Statistic	Number of Items	Easy Scale	Difficult Scale	Difference
Mean	12	49.49	37.93	11.58
Std. Dev.	12	7.31	9.16	6.12

LOGOG Attitude Estimates[b]				
Statistic	Number of Items	Easy Scale	Difficult Scale	Standardized Difference
Mean	12	.07	.01	.05
Std. Dev.	12	.98	1.15	1.48

Note. The correlation between the LOGOG attitude estimates for the easy and hard subscales was r = .71.

[a]The sample size was 490 respondents since one person with the maximum score on all items was deleted from analysis.

[b]The sample size was 321 respondents due to the deletion of 170 nonconvergence cases of attitude estimation when using maximum likelihood.

Table 4

Comparison of Person Attitude Estimates[a] for Easy and Hard Subscales for the Three Data Sets

ADCOM Scale				
Statistic	Number of Items	Easy Subscale	Difficult Subscale	Difference
Mean	12	-.04(475)	-.07(475)	.04(475)
Std. Dev.	12	1.08(475)	1.09(475)	.63(475)
SD-U Scale				
Statistic	Number of Items	Easy Subscale	Difficult Subscale	Difference
Mean	12	-.18(900)	.16(912)	.05(817)
Std. Dev.	12	.99(900)	1.01(912)	.33(817)
SD-DS Scale				
Statistic	Number of Items	Easy Subscale	Difficult Subscale	Difference
Mean	12	-.03(972)	.03(974)	.05(948)
Std. Dev.	12	1.11(972)	1.10(974)	.48(948)

Note. The parentheses indicate the number of respondents upon which the statistics were based. Cases of nonconvergence of trait estimation were deleted.

[a]The person attitude estimates were obtained from the empirical maximum likelihood program using item parameters estimated with Samejima's method.

REFERENCES

Andersen, E.B. Sufficient statistics and latent trait models. Psychometrika, 1977, 42, 69-81.

Andrich, D. Application of a psychometric rating model to ordered categories which are scored with successive integers. Applied Psychological Measurement, 1978, 2, 581-594.

Bock, R.D. Personal communication, April 2, 1980.

Kolakowski, D., & Bock, R.D. Maximum likelihood item analysis and test scoring: logistic model for multiple item responses. Ann Arbor, Mich.: National Educational Resources, Inc., 1973.

Lord, F.M. Applications of item response theory to practical testing problems. Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publ., 1980.

Lord, F.M. & Novick, M.R. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

Rasch, G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Educational Research, 1960.

Reckase, M.D. An interactive computer program for tailored testing based on the one-parameter logistic model. Behavior Research Methods & Instrumentation, 1974, 6, 208-212.

Samejima, F. Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph Supplement, 1969, No. 17.

Valentine, J.W. Audit of administrator communication. Columbia, Mo.: copyright, 1978, by Jerry W. Valentine.

Wherry, Sr., R.J., Naylor, J.C., Wherry, Jr., R.J., and Fallis, R.F. Generating multiple samples of multivariate data with arbitrary population parameters. Psychometrika, 1965, 30, 303-313.

Whitely, S.E., & Dawis, R.V. The nature of objectivity with the Rasch model. Journal of Educational Measurement, 1974, 11, 163-178.

Wright, B.D. Sample free test calibration and person measurement. Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton, New Jersey: Educational Testing Service, 1968, 85-101.